

Feasibility of Local and Statewide Assessments in an A-F Accountability System

Dr. Thomas M. Haladyna

Arizona State University

January 11, 2016

¹ This essay addresses the related topics of the *test-based accountability* systems, such as the one used in Arizona, and using statistical methods to analyze test scores in an accountability system. The term *test-based accountability* is also known as *value-added measurement* (VAM).

1. Limitations of VAM in the current accountability system are presented and discussed.
2. Using locally produced tests and various data analysis strategies proposed for such an accountability system need validity.

Test-based Accountability

In any educational accountability system, we have many concepts and principles that guide us. A good understanding of these concepts and principles equips us to make better decisions about how to improve student learning.

First, student achievement is very complex. Achievement includes many abilities: reading, writing, speaking, listening, mathematical and scientific problem solving, critical thinking, citizenship and responsibility, motivation, and attitude. An ability like reading is a complex mental structure consisting of knowledge and skills and the use of knowledge and skills in complex ways. It takes many years to develop each ability. Ability grows very slowly. Because of this complexity, measuring an ability, such as reading, is very challenging.

Second, the most potent influence on student learning is time. The more a student is engaged in learning relevant and important content, the more they learn. Teachers who can engage students in learning activities a high percentage of the time allowed are the most effective.

Third, the measurement of each student's ability requires multiple, validated measures. No single test or even short battery of tests produces validly interpreted results. Our national standards for testing and for the use of test scores recommend multiple, highly correlated measures. If a set of test scores is to be used, it should be validated. Validation is a demanding enterprise. Typically, test score use of this kind is seldom validated.

Fourth, the concept of accountability appears to be focused mainly on schools and teachers. As we know in Arizona, school districts are unequally funded and supported. Teachers often lack resources to serve a student population that is more than 50% at risk. An effective accountability system would involve many constituencies: the state board of education, legislators, the governor, teacher educators in the state, school district leaders, school board members, principals, teachers, students and their parents. Such a system has yet to be devised.

¹Information about my qualifications for writing this essay can be found in Appendix A

Fifth, we have various sources that guide us in the development of a test and the interpretation and use of test scores. These are listed in the annotated bibliography at the end of this essay. Briefly, the credibility of any accountability system rests on validity—the seeking of truth for any test score interpretation or use. Validity is essential for any test score interpretation or use.

Sixth, we have known for more than 60 years, that a major determinant of student achievement is social capital, which refers to family, home, and environmental influences (Coleman, 1988). For instance, parental education and income are powerful predictors of student achievement. This factor is so dominant that no other factor comes close to representing that strong of an influence. If a student lacks social capital, the student is at-risk. These at-risk students live in poverty, may not speak English fluently, may have a physical, mental, or emotional disability, or live in isolation. Our state has many students who are classified at-risk. Some students have several or all of these limitations that place them at-risk. These students require intensive services that our state does not currently adequately provide. This lack explains why Arizona is ranked so low nationally in education.

Seventh, a recent book by Dr. Audrey Beardsley of ASU (*Rethinking value-added models in education* (2014) is the most comprehensive, compelling analysis of test-score accountability to date. Dr. Beardsley is a leading authority with a national reputation. Her research and voice have been influential in the increasing scrutiny of teacher-based accountability systems. As a result of her efforts, states and school districts have been more careful about the use of test scores, particularly when it comes to making personnel decisions regarding teachers.

What Does the American Statistical Association Say About VAM?

This prestigious organization has many admonitions about VAM. Appendix B contains 10 major points. A review of these points reveals the responsibility for using tests scores wisely and fairly is considerable. The complexity of such test score use is difficult because of many threats to validity.

What Does the American Educational Research Association (AERA) Say About VAM?

AERA was founded in 1916 and continues to be the largest organization in the world devoted to the study and improvement of student learning. AERA has also issued a statement about VAM. Excerpts of this statement are listed in Appendix C.

Summary

Measuring student achievement validly requires great care. Using test scores to evaluate student learning and then attribute these results to teachers, schools, and school districts is risky. If the intent is to improve student learning, a system of accountability should be comprehensive and thoughtfully planned, designed, and implemented.

Multi-level, Multiple Measures, and/or Data Points

Statisticians may use *fireworks* statistical methods that make promises that are not supportable by reasoning or critical thinking. There is a tendency to accept statistical methods without understanding the implications of their use. Statisticians may develop methods of analysis that ignore fundamental concerns like validity. That is why peer review is so important.

A fundamental quest in student achievement testing is measuring growth. That requires tests that are layered to the extent that we can measure student progress from grades three through eight. Measurement before grade three has many limitations to validity. After grade eight, the curriculum of each student is varied to the extent that measuring basic abilities like reading, writing, and mathematical problem solving become moot.

We have outstanding experts at multi-level measurement. Some good references are Kolen and Brennan (2014) and the *Handbook of Test Development* (Downing and Haladyna, 2006; Lane, Raymond, and Haladyna, 2015). Providing a vertical scale consisting of several tests is challenging because the content of what is measured (for example, reading) changes from grade to grade. Nonetheless, it has been done. The danger of multi-level measurement is that is the accountability system comprehensive enough to encompass the entire curriculum or just choose several topics (typically, reading, writing, and mathematics). As we know, the choice of just three student abilities compels teachers to teach only what is tested because that is what counts.

The practice of using multiple measures is highly recommended by the AERA (2002). Test developers and testing experts recommend that important abilities to be measured should be done with highly validated, correlated measures. This improves reliability and validity. However, multiple measures increases the cost of testing and takes more student time that can be devoted to learning so testing must be done prudently and wisely. The other challenge is that multiple measures should be done for the whole of the curriculum not several aspects of the curriculum.

Locally developed, supplemental measures of student achievement may not be held to the same high standards desired. This lack of validation puts the state of Arizona at risk if high-stakes decisions are made with results from such local tests.

One of the most vexing issues in accountability systems is understanding the variation of test scores and the related interpretation of cause-and-effect. VAM accountability focuses on variations in test scores of classroom or school data. One caveat is effect size—which is the practical difference of this variation. Often, the effect size is near zero when the result is statistically significant. Statistical significance is a necessary but not sufficient condition for drawing a conclusion. This would lead us to interpret a difference as an effect when, in fact, the effect is negligible. For example, in a weight loss study, it was found that a treatment produced a statistically significant weight loss of three ounces for 1,000 patients. Although

statistically significant; practically, three ounces is very close to zero. Accountability systems seldom take into account effect size.

The second aspect of this vexing issue is cause-and-effect. Attributing gains in student achievement or lack of a gain to a teacher has so many logical and statistical difficulties that it can and has been legally challenged. One of the biggest problems with cause-and-effect is cheating on the test (see Haladyna and Downing, 2004; Haladyna, Nolen, & Haas, 1989). When school leaders and teachers are forced to be accountable to reading, writing, and mathematics test scores, they teach to the test or cheat. Thus, fraud takes the place of effective educational programs when an insufficient accountability system is in place.

Recommendations

1. The Arizona Board of Education should consult with experts in testing who understand the pitfalls and challenges of accountability and the use of test scores for evaluating student performance. The state has many such experts. To name a few: Dr. Joe Ryan, former dean and faculty member at Arizona State University, Dr. Audrey Beardsley, Associate Professor at Arizona State University, Dr. Joe O'Reilly of the Mesa School District, Dr. Ed Sloat, who has served in many capacities in our state as early back as 1986, and Dr. Joan Jameson of Northern Arizona University, who has been one of the innovators in test score validation.
2. A good way to accomplish this is to establish a technical advisory committee, as the state does for its student testing. Dr. Ryan and I have served on such committees in Arizona and other states. These experts provide credible advice to state and other testing agencies regarding valid use of test scores.
3. We have excellent resources that inform us about the wisdom and effectiveness and guide us in test development and use of test scorers. This is a bibliography that should be included. For example, the *Handbook of Test Development* (2006, 2015) is one source. The *Standards for Educational and Psychological Testing* (AERA, 2014) is another source. The annotated bibliography is in the appendix. Such sources and the principles represented within should be observed.
4. Regarding accountability, if some exotic statistical methods are adopted, peer review is very advisable. Law suits have arisen up due to invalid use of test scores to terminate teachers, fail students, or close schools. If methods are not validated, states will lose in the courts.

References

American Educational Research Association (AERA) (2000). *Position statement on high-stakes testing*. Source:

<http://www.aera.net/AboutAERA/AERARulesPolicies/AERAPolicyStatements/PositionStatementonHighStakesTesting/tabid/11083/Default.aspx>. *AERA is the largest scientific organization dedicated to the advancement of education in the world. Its statement provides important guidelines that we follow in the development student achievement tests and the interpretation and use of test scores.*

American Educational Research Association, American Psychological Association & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association. *The latest version of an authoritative guide for test developers and users. Contains many guidelines and explains concepts, principles, and procedures useful to test developers and users.*

Amrein-Beardsley, A. (2014). *Rethinking value-added models in education*. NY: Routledge. *This recent book increases scrutiny on the unfair and invalid use of test scores for evaluating teachers. It has been well received in the educational community, and Dr. Beardsley of ASU has received many accolades and recognition for her work.*

Coleman, J. S. (1988). Social capital in the creation of human capital. *American Journal of Sociology*, Vol. 94, *Supplement: Organizations and Institutions: Sociological and Economic Approaches to the Analysis of Social Structure* (1988), pp. S95-S120. *This famous article has been cited more than 31,000 times. With certainty, social capital is a major cause of student achievement and other desirable outcomes of schooling.*

Downing, S. M., & Haladyna, T. M. (Eds.) (2006) *Handbook of test development* (1st ed.). NY: Routledge. *This popular volume contains a compendium of topics written by leading experts in the field. The book has been very popularly received by the educational testing community as a resource in developing tests and validating test score interpretations and uses.*

Haladyna, T. M., & Downing, S. M. (2004). *Construct-irrelevant variance in high-stakes testing*. *Educational Measurement: Issues and Practice*, 23(1), 17–27. *This often cited article has made users of tests more aware of threats to validity that come from various sources. There is no doubt that test scores may be biased to these many threats. Some of these threats arise from invalid uses of test scores, such as for accountability.*

- Haladyna, T. M., Nolen, S. B., & Haas, N. S. (1991). Raising standardized achievement test scores and the origins of test score pollution. *Educational Researcher*, 20(5), 2-7. *This well-cited study was funded by the Arizona legislature in the 1980s to investigate how teachers use standardized test scores. The findings from self-confessed teachers is that about 11% cheat on these tests during an era where test scores did NOT have high-stakes consequences.*
- Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). Westport, CT: American Council on Education and Praeger. *From the leading writer on validity, this chapter discusses the substance and intricacies of validity and validation for modern testing purposes.*
- Kolen, M. J., & Brennan, R. L. (2014). *Test equating, scaling: Methods and practices* (3rd ed.). New York, NY: Springer. *This popular third edition is comprehensive in its treatment of equating tests, comparable scaling, and multi-level testing.*
- Lane, S., Raymond, M. R., Haladyna, T. M. (Eds.) (2015). *Handbook of test development* (2nd ed). NY: Routledge. *This second edition complements the first edition and introduces new topics that update validity and link to the new Standards for Educational and Psychological Testing.*

Appendix A

Dr. Thomas Haladyna has served in the field of education for more than 50 years. He was an elementary school teacher, faculty member at several universities, a research professor, test director at ACT, visiting scholar at the Naval Personnel Research and Development Center, visiting NAEP scholar at the Educational Testing Service (ETS), and consultant to more than 100 clients desiring services or improvements in testing programs. He has authored, co-authored, and edited 14 books and authored several hundred journal articles, white papers, research reports, test evaluations, opinions, and technical reports.

Appendix B

Admonitions from the American Statistical Association About Value-added Measurement

1. VAMs are complex statistical models, and high-level statistical expertise is needed to develop the models and [emphasis added] interpret their results.
2. Estimates from VAMs should always be accompanied by measures of precision and a discussion of the assumptions and possible limitations of the model. These limitations are particularly relevant if VAMs are used for high-stakes purposes.
3. VAMs are generally based on standardized test scores, and do not directly measure potential teacher contributions toward other student outcomes.
4. VAMs typically measure correlation, not causation: Effects – positive or negative – attributed to a teacher may actually be caused by other factors that are not captured in the model.
5. Under some conditions, VAM scores and rankings can change substantially when a different model or test is used, and a thorough analysis should be undertaken to evaluate the sensitivity of estimates to different models.
6. VAMs should be viewed within the context of quality improvement, which distinguishes aspects of quality that can be attributed to the system from those that can be attributed to individual teachers, teacher preparation programs, or schools.
7. Most VAM studies find that teachers account for about 1% to 14% of the variability in test scores, and that the majority of opportunities for quality improvement are found in the system-level conditions. Ranking teachers by their VAM scores can have unintended consequences that reduce quality.
8. Attaching too much importance to a single item of quantitative information is counter-productive—in fact, it can be detrimental to the goal of improving quality.
9. When used appropriately, VAMs may provide quantitative information that is relevant for improving education processes...[but only if used for descriptive/description purposes]. Otherwise, using VAM scores to improve education requires that they provide meaningful information about a teacher's ability to promote student learning...[and they just do not do this at this point, as there is no research evidence to support this ideal].
10. A decision to use VAMs for teacher evaluations might change the way the tests are viewed and lead to changes in the school environment. For example, more classroom

time might be spent on test preparation and on specific content from the test at the exclusion of content that may lead to better long-term learning gains or motivation for students. Certain schools may be hard to staff if there is a perception that it is harder for teachers to achieve good VAM scores when working in them. Over reliance on VAM scores may foster a competitive environment, discouraging collaboration and efforts to improve the educational system as a whole.

Source: <http://vamboozled.com/american-statistical-association-asa-position-statement-on-vams/>

Appendix C

Statements from the AERA Regarding VAM

Here are AERA's eight technical requirements for the use of VAM:

1. "VAM scores must only be derived from students' scores on assessments that meet professional standards of reliability and validity for the purpose to be served...Relevant evidence should be reported in the documentation supporting the claims and proposed uses of VAM results, including evidence that the tests used are a valid measure of growth [emphasis added] by measuring the actual subject matter being taught and the full range of student achievement represented in teachers' classrooms" (p. 3).
2. "VAM scores must be accompanied by separate lines of evidence of reliability and validity that support each [and every] claim and interpretative argument" (p. 3).
3. "VAM scores must be based on multiple years of data from sufficient numbers of students...[Related,] VAM scores should always be accompanied by estimates of uncertainty to guard against [simplistic] overinterpretation[s] of [simple] differences" (p. 3).
4. "VAM scores must only be calculated from scores on tests that are comparable over time...[In addition,] VAM scores should generally not be employed across transitions [to new, albeit different tests over time]" (AERA Council, 2015, p. 3).
5. "VAM scores must not be calculated in grades or for subjects where there are not standardized assessments that are accompanied by evidence of their reliability and validity...When standardized assessment data are not available across all grades (K–12) and subjects (e.g., health, social studies) in a state or district, alternative measures (e.g., locally developed assessments, proxy measures, observational ratings) are often employed in those grades and subjects to implement VAM. Such alternative assessments should not be used unless they are accompanied by evidence of reliability and validity as required by the AERA, APA, and NCME *Standards for Educational and Psychological Testing*" (p. 3).
6. "VAM scores must never be used alone or in isolation in educator or program evaluation systems...Other measures of practice and student outcomes should always be integrated into judgments about overall teacher effectiveness" (p. 3).
7. "Evaluation systems using VAM must include ongoing monitoring for technical quality and validity of use...Ongoing monitoring is essential to any educator evaluation program and especially important for those incorporating indicators based on VAM that have only

recently been employed widely. If authorizing bodies mandate the use of VAM, they, together with the organizations that implement and report results, are responsible for conducting the ongoing evaluation of both intended and unintended consequences. The monitoring should be of sufficient scope and extent to provide evidence to document the technical quality of the VAM application and the validity of its use within a given evaluation system” (AERA Council, 2015, p. 3).

8. “Evaluation reports and determinations based on VAM must include statistical estimates of error associated with student growth measures and any ratings or measures derived from them...There should be transparency with respect to VAM uses and the overall evaluation systems in which they are embedded. Reporting should include the rationale and methods used to estimate error and the precision associated with different VAM scores. Also, their reliability from year to year and course to course should be reported. Additionally, when cut scores or performance levels are established for the purpose of evaluative decisions, the methods used, as well as estimates of classification accuracy, should be documented and reported. Justification should [also] be provided for the inclusion of each indicator and the weight accorded to it in the evaluation process...Dissemination should [also] include accessible formats that are widely available to the public, as well as to professionals” (p. 3-4).

As per the conclusion: “The standards of practice in statistics and testing set a high technical bar for properly aggregating student assessment results for any purpose, especially those related to drawing inferences about teacher, school leader, or educator preparation program effectiveness” (p. 4). Accordingly, the AERA Council recommends that VAMs “not be used without sufficient evidence that this technical bar has been met in ways that support all claims, interpretative arguments, and uses (e.g., rankings, classification decisions)” (p. 4).

AERA Council. (2015). AERA statement on use of value-added models (VAM) for the evaluation of educators and educator preparation programs. *Educational Researcher*, X(Y), 1-5. doi:10.3102/0013189X15618385 Retrieved from <http://edr.sagepub.com/content/early/2015/11/10/0013189X15618385.full.pdf+html>